图书情报工作
第 62 卷 第 10 期 2018 年 5 月

基于 LDA 主题模型的学术谱系内知识传承研究*

——以谈家桢为核心的遗传学学术谱系为例

■ 刘俊婉 杨波 王菲菲 徐硕
北京工业大学经济与管理学院 北京 100124

摘要: [目的/意义]学术谱系以知识传承的方式助推科学发展。研究知识传承特征,探索学术谱系传承模式及其学术产出影响效能,为探索人才成长规律及人才政策制定提供参考。[方法/过程]基于 LDA 主题模型,以遗传学领域谈家桢为核心的学术谱系成员发表的期刊文献为研究对象,对该谱系成员的研究主题进行抽取,借鉴生物学"遗传"与"变异"的概念,根据主题相似度将谱系成员划分为"遗传学者""变异学者"和"非遗传非变异学者",并对三种学者的学术绩效进行分析。[结果/结论]分析结果表明,谈家桢学术谱系内"遗传学者"和"变异学者"的学术绩效相对较高;"非遗传非变异学者"的数量占比最多,但学术绩效相对较低;"遗传学者"与"变异学者"在不同主题下的分布具有明显差异。

关键词: 学术谱系 知识传承 主题模型 遗传学
分类号: G250
DOI: 10.13266/j.issn.0252-3116.2018.10.011

1 引言

"吾生也有涯,而知也无涯",在世界无穷奥秘面前个人的力量是渺小的。但涓流可至沧溟水,知识在人类繁衍生息中不断积累,通过在不同谱系、流派中代代传承,铸就当今繁盛的科学体系。中国科学技术的发展历史,具体到每一个研究领域,是以各学科带头人所创立的学术谱系的建立、拓展和衍生的过程。学术谱系是学术共同体自觉认同的范式在时间上的延续和传递,它是构成学术积累的必要条件[1]。学术谱系内关系的确认是通过知识在老师与学生之间由老师输送至学生的事实为基础建立的。学术谱系研究具有重要的学术价值,它突破了以往科学史研究的边界,涉及由学术谱系传承过程中数代科学家所构成的庞大的研究群体,在时段上考察历时达数十年乃至近百年的学术谱系的发生发展过程。梳理学科谱系关系,厘清知识传承脉络,探究不同知识传承模式对学术产出的影响效能,对于探究科技人才成长规律以及科技政策制定具有重要参考价值。

科学研究领域纷繁多样,不同研究领域的学术谱系、同一研究领域中各学术谱系之间的传承方式都有着或多或少的区别,但知识自上而下的流动方向却是亘古不变的。在知识的流动过程中作为接收方的学生对信息的积累与运用不同,导致科研绩效也会存在差异,这两者之间是否具有相关性?早期学术谱系的研究通常致力于寻找某些成功学者的成长历程与学术起源[2],这些研究在定性基础上阐述学者对研究对象所获成就的个人观点,由于缺乏足够的数据支撑,并未得出普适的人才发展规律。随着"h 指数"等评价指标以及诸如 Academic Tree 等学术谱系数据库的出现,定量化的学术谱系分析成为可能[3]。至此,部分学者开始利用科研文献作为切入点展开与学术谱系评价有关的量化研究。由于学术谱系量化研究出现较晚,发展尚不成熟,目前的学术谱系量化研究主要集中在通过学术文献的合作与引用关系,建立学者之间的关系网络,进而在系统、团体、个人三个关系层面展开研究。例如:R. D. Malmgren 等通过计量学者教职生涯中的学生数量以及学生的学术影响力发现,学者在学术生涯

* 本文系国家自然科学基金青年项目"共生视角下的院士科学合作网络结构与演化趋势研究:以中美两国科学院院士为例"(项目编号:71603015)和北京市自然科学基金项目"基于技术共生网络结构探测和演化的新兴趋势识别研究"(项目编号:9182001)研究成果之一。
作者简介:刘俊婉(ORCID:0000-0001-7911-4681),副教授,博士,E-mail:liujunwan@bjut.edu.cn;杨波(ORCID:0000-0003-2609-3885),硕士研究生;王菲菲(ORCID:0000-0002-1717-9719),副教授,博士;徐硕(ORCID:0000-0002-8602-1819),副研究员,博士。
收稿日期:2017-11-15 修回日期:2018-01-24 本文起止页码:76-84 本文责任编辑:王善军

的前2/3时期相较于后1/3时期具有更强的学术繁衍能力[4]。国内学者运用社会网络分析方法对第四纪学术谱系的合作网络特征以及学术传承分析显示，处于学术大本营的学生与导师的合作具备长期稳定的特征，合作强度明显高于离开学术大本营到其他单位工作的学生[5]。C. Sugimoto等通过建立谱系树的方法，分析谱系成员研究领域的转移特征并将其用于跨学科理论的研究。上述研究对于学术谱系成员的关系网络结构等外部特征研究较多，但还未对谱系内以文本为载体的思想、知识的传承进行量化。因此，笔者选取中国近代遗传学领域以谈家桢为核心的学术谱系为研究对象，通过主题模型的方法，对谱系内师生之间以及学生毕业前后研究主题的变化进行研究，探索谱系发展过程中谱系成员研究主题的变化程度，以及研究主题变化程度与其科研绩效的关系，以此探索谱系内知识传承的规律性特征，探寻科技人才的成长路径，从而为科研管理等相关部门制定科技人才政策提供参考。

2 研究方法与数据获取

2.1 学术谱系知识传承研究方法与技术路线

本研究利用师生关系数据构建学术谱系树，并从谱系内学者的论文标题、摘要和关键词抽取论文特征词，进而根据论文与学者的对应关系构建学者特征词库。通过十折交叉验证的方法获得LDA(Latent Dirichlet Allocation)[6]主题模型的最佳主题数量。在此基础上通过LDA主题模型获取每一位学者在固定维度上的主题分布向量，使用JS(Jensen-Shannon)距离方法获得不同学者对应向量之间的距离，进而获得两学者之间的主题相似度并依此将学者划分为“遗传学者”“非遗传非变异学者”和“变异学者”。最后，对分属不同类型学者的学术绩效和总体主题分布进行了特征分析。具体研究路线如图1所示，分为4个部分：数据获取、数据预处理、主题抽取与相似度计算、特征分析。

数据获取分为两部分：谱系数据获取与文献数据获取。学术谱系的研究基础需要在广泛扎实收集史料、确定谱系代际关系的基础上识别并绘制出学术谱系树，该步骤的关键是要确保数据的真实性。而期刊文献是学者阶段性研究成果的主要载体，是科学知识在学术共同体中更新、传播和交流的主要形式，是学者的重要学术产出。为了对谱系成员的知识传承进行研究，将谱系内学者发表的所有中文期刊文献及其研究生学位毕业论文作为研究对象。

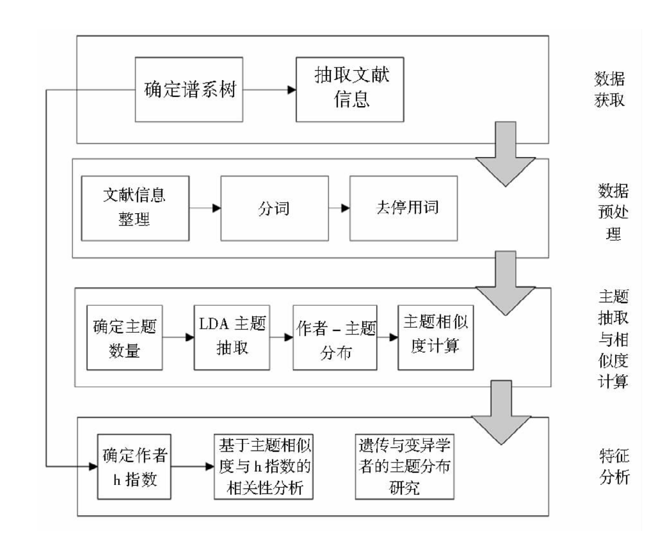


图1 学术谱系知识传承研究路线图

数据预处理主要涉及自然语言预处理工作，LDA主题模型的数据输入是以特征词为代表的文档集合。本模型输入数据样本分为两种：①集合中的每一个文档对应一位学者的全部论文信息，此样本用来分析学者与其导师研究方向的差异性；②集合中的每一个文档对应一位学者毕业前的全部论文信息或毕业后的全部论文信息，用来分析学者毕业前后研究方向的差异性。两个样本中各文档间不具有先后顺序差别，学者研究方向的不同在文本层次中体现为特征词集合的差异。数据预处理主要包含对学者文献的文本分词以及去停用词两部分操作。在文本分词过程中需要在特征词过度切割及切词粒度过小中合理进行取舍。去停用词的过程中排除语气助词、副词、介词、连接词等噪音词汇的干扰。

通过LDA主题模型算法获取每位学者对应的研究主题分布向量，在此基础上通过向量相似度可以定量描述老师与学生、学生毕业前后研究方向的变化、知识传承的程度。LDA作为一种非监督的机器学习方法，可以用来识别大规模文档集或语料库中潜藏的主题信息。为达到本文的研究目的，论文将LDA模型中的文档层替换为以谱系学者所发表的论文集为实体的作者层，每一个作者代表其论文集的主题、关键词、摘要的集合，此转换实际与AT模型(Author Topic Model)[7]原理一致。模型假设作者是若干主题的混合分布(Author-Topic)，而主题又是关于单词的概率分布(Topic-Word)。这种假设使得作者数据集被投影到主题空间，从而降低了大规模数据处理过程的时间复杂度[8]。其中，超参数α、超参数β、作者主题分布$\boldsymbol{\theta}$、主题-词汇分布φ均为隐含变量。包含M个作者

的作者集 $D=\{d_1,d_2,d_3,\cdots,d_M\}$，这些作者的研究分布于 K 个主题 $Z=\{z_1,z_2,z_3,\cdots,z_K\}$。作者文本集中的所有特征词构成了一个词汇表 $W=\{w_1,w_2,w_3,\cdots,w_N\}$。则每个作者所对应的概率密度函数如公式(1)所示：

$$P(w \mid \alpha,\beta)=\int P(\boldsymbol{\theta} \mid \alpha)\left(\prod_{n=1}^{N}\sum P(z_n \mid \boldsymbol{\theta})P(w_n \mid z_n),\beta\right)d\boldsymbol{\theta} \quad \text{公式(1)}$$

上述方法原理与 AT 模型一致，但没有直接使用该模型进行主题抽取，原因在于：AT 模型支持对一篇论文中的多作者研究主题分布的测度，但并没有考虑不同署名顺序的学者对论文的贡献是不同的，因此 AT 模型存在一定的局限性，最终的测度结果会产生一定的偏差。基于此，笔者在 LDA 模型的基础上分析作者的研究主题分布。

应用 LDA 进行主题建模时，主题个数是由建模者指定的。确定语料库的最优主题个数是构建主题模型时必须考虑的重要问题之一。本研究采用统计语言模型中常用的评价指标即困惑度（Perplexity）来评价模型性能，确定最佳主题个数。困惑度分析可以表征生成模型的质量，本研究采用十折交叉验证（将学者文献数据集随机分成 10 等份，依次使用其中 9 份数据作为训练模型，另外的 1 份数据作为测试模型）的方式进行验证，将 10 次困惑度分析结果的平均值作为最终结果困惑度的数值。其实际代表文档集中包含句子相似性几何均值的倒数，越低的困惑度数值表示模型的效果越好[9]。对于测试集的 M 个作者，困惑度定义如公式(2)所示：

$$perplexity(D_{test})=exp\left(-\frac{\sum_{m=1}^{M} ln(P(w_m))}{\sum_{m=1}^{M} N_m}\right) \quad \text{公式(2)}$$

其中，D_{test} 为作者测试集，即 6 623 篇作者文献的集合，N_m 代表第 m 个作者语料中的单词数目，$P(w_m)$ 代表 LDA 产生该作者文本集的概率，如公式(3)所示：

$$P(w_m)=\prod_{i=1}^{N_m}\sum_{k=1}^{K}p(w_i \mid z_i=k)p(z_i=k \mid w_m) \quad \text{公式(3)}$$

X. Wang 和 A. Mccallum 通过实证分析验证了 JS 距离相对欧式距离、余弦距离等在学者主题向量的区分度方面更具优势[10]，因此笔者采用 JS 距离来衡量学者之间主题的相似度。设 Θ 为作者集 D 在主题集 Z 上的全体离散概率分布，则对任意 $P,Q\in\Theta$，JS 距离的计算如公式(4)所示：

$$J(P \mid Q)=\frac{1}{2}[F(P \mid Q)+F(Q \mid P)]$$
$$=\frac{1}{2}\sum_{m=1}^{M}\left(p_m\, ln\frac{2p_m}{p_m+q_m}+q_m\, ln\frac{2q_m}{p_m+q_m}\right) \quad \text{公式(4)}$$

学者主题相似度 $Sim(P,Q)$ 计算如公式(5)所示，其中 P,Q 分别为两学者的主题分布向量。

$$Sim(P,Q)=1-JS(P \mid Q) \quad \text{公式(5)}$$

特征分析分为两部分：①学者主题相似度与学者产出绩效的相关性分析。通过学术谱系树中的师生关系数据与已获取的主题相似度集合，计算老师与学生的研究主题相似度和学生毕业前后的研究主题相似度。学者主题相似度旨在对两个学者研究方向与研究领域的一致性——知识传承程度进行度量，并结合生物学“遗传”与“变异”概念，将学者划分为“遗传学者”“非遗传非变异学者”以及“变异学者”。通过已有文献集合数据获取每位学者的 h 指数作为学者的产出绩效测度指标。在此基础上分析知识“遗传”、知识“变异”与学者学术绩效的相关性。②遗传与变异学者的主题分布研究。绘制三类学者的研究主题分布图，比较不同类别的学者在不同主题上分布的差异性。同时，对整个谱系研究主题的分布做一个宏观的分析。

2.2 数据获取

本项研究的学术谱系数据来源于中国科协“当代中国科学家学术谱系研究”课题的项目成果，该项目成果对遗传学、医学、化学、物理以及农学 5 个领域的学术谱系进行了系统梳理，从中可以找到清晰的遗传学谱系成员脉络和代际关系，笔者选取遗传学领域中发展历史最久、史料最全、规模最大的谈家桢学术谱系作为研究对象。

谈家桢，1932 年毕业于北京燕京大学研究院并获硕士学位，1936 年获得美国加州理工学院博士学位，50 年代建立了中国第一个遗传学专业，1999－2008 年任教浙江大学生物学系。谈家桢先生 60 年的教学生涯中为中国遗传学领域培养了众多学术精英，例如在作物遗传育种上有突出贡献的季道藩、汪丽泉、唐绝等 30 余人。如今，谈家桢遗传学谱系已有前后 5 代成员，能够获得详细师承信息的就有 532 位。本研究获得了谈家桢学术谱系内 532 名学者的姓名、学位、获取学位时间、获取学位所在院校、导师姓名等有效信息。笔者定义师生之间的关系为代际间关系，其三代谱系树见图 2。

在确定谱系树的基础上，本项研究获取谱系内学者所有发表的期刊文献信息，该数据从中国知网（CNKI）

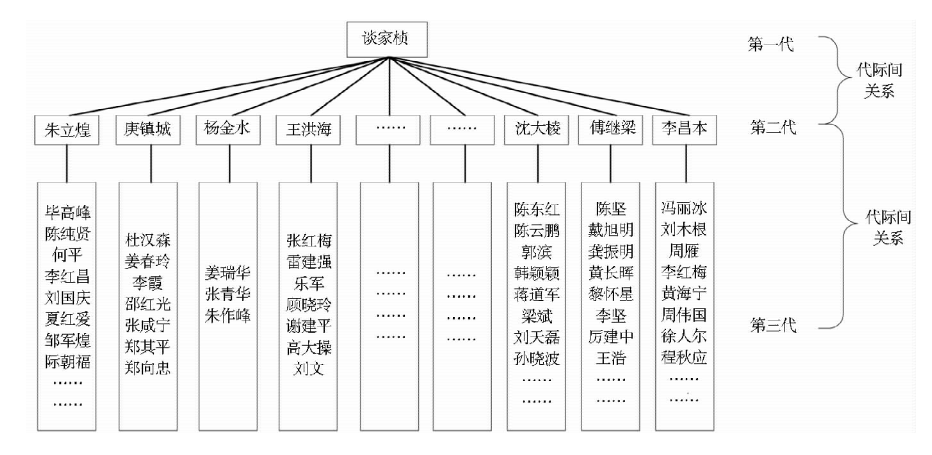


图 2　谈家桢学术谱系树

逐一检索获得。鉴于谱系成员数量庞大，本研究选取了以谈家桢作为第一代谱系成员延续至第三代的共计 234 名谱系成员作为研究对象。在中国知网(CNKI)上以 234 名学者的姓名、发表文献时间、主题词、作者机构等信息作为检索要素，以(SU =‘遗传’+‘基因’+‘DNA’+‘染色体’+‘甲基’+‘突变’+‘RNA’+‘烷基’+‘变异’+‘启动子’) AND (AU =‘作者姓名’) AND (AF =‘作者机构’) AND (YE BETWEEN (‘1960’,‘2017’))作为检索式进行文献检索，同时将学者的导师姓名、获取学位的时间作为判别因素辅以人工手动甄别以消除重名带来虚假信息的影响，共计获得了 6 623 篇期刊文献数据(包括文献的题目、关键词、摘要、发表期刊、被引频次等)。需要说明的是，研究伊始，希望对谱系内成员发表的 SCI 论文进行分析，在 Web Of Science(WOS)数据库中以((SU = genetics & hereditism) AND 语种:(English) AND 文献类型:(Article))作为高级检索式进行检索，共获得 463 214 篇遗传学领域文献，文献数量庞大。鉴于中国学者姓名在 SCI 数据库中的重名现象严重，且 2006 年之前的 SCI 论文没有中国作者的姓名全称，再加上作者所在机构名称形式多变，因此姓名消歧工作进展困难，精准获取作者发表 SCI 论文的全面信息比较困难，同时学术谱系成员在 CNKI 数据库发表的论文信息能够满足本项研究对数据样本的要求。鉴于上述原因，本研究选用谱系成员发表的 CNKI 论文集作为研究对象。

3　以谈家桢为核心的遗传学学术谱系知识传承分析

3.1　以谈家桢为核心的学术谱系 LDA 主题分析

首先，针对文献数据集的特征，自编 Python 脚本，结合 jieba 中文分词、jieba 词性标注进行数据预处理操作，在去除语气助词、副词、介词、连接词的基础上最终共获得 234 位学者的 46 万多个特征词。

采用交叉验证的方法评估模型的性能，在主题数量 K 分别为 5、10、15、20、25、30 的情况下进行试验，得到不同情况下的困惑度。实验结果如图 3 所示，当主题数为 20 的情况下困惑度达到最低值，因此谈家桢遗传谱系 LDA 主题建模的最优主题个数选为 20。

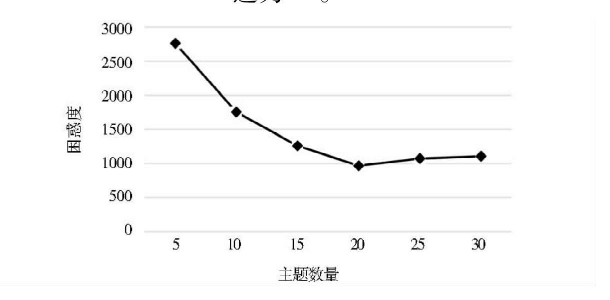


图 3　LDA 主题数量与困惑度分布图

基于预处理后的数据集，应用 LDA 主题模型方法，对谈家桢遗传学领域的文献数据进行主题分析，得到主题－词分布。由于篇幅有限，仅列出所获得 20 个主题中具有代表性的 5 个主题以及分别在 5 个主题中权重最高的前 15 个单词，如表 1 所示。通过充分了解遗传学领域的研究背景、咨询领域专家，结合该科研领域研究生导师的研究方向等信息，确定了研究主题的内容，并对主题内容进行“描述”。根据学者－主题分布，表 2 为谱系内三位第二代学者的研究主题分布，从向量数值可以观察出这三位学者的主要研究方向。例如朱立煌在“主题 15”(植物育种遗传学)中分布值最大，通过资料调查得知朱立煌的研究方向是水稻分子遗传学和基因组研究；曾益滔在“主题 1”(血液遗传学)中分布值最大，通过资料调查得知其在血红蛋白病领域的珠蛋白化学、基因结构和功能，以及地中海贫血基因治疗等方面的研究成绩卓著；冯蜀举在“主题 7”(基因组学)中分布值很大，调查资料得知其师从施立明，主要是致力于研究真核细胞染色体结构与功能、细胞分类学和核型进化。

这里选取弟子数量最多的第二代代表性学者“朱立煌”为例，分析其与子代学生之间的主题相似度。表 3 列出了朱立煌与子代学生主题相似度前 5 位与后 5 位的数据。对全部数据进行计算得到谱系内成员的相似度处于[0.487,0.747]区间内。

表 1　谈家桢学术谱系主题－词分布

第 1 类:血液遗传学	血红蛋白、贫血、地中海、珠蛋白、Hb(血红蛋白)、小鼠、转基因、患者、链、胚胎、DNA、产前诊断、胎儿、家系、性别
第 4 类:植物遗传学	稻、杂草、同工酶、群落、土壤、多样性、酯酶、中国、农田、物种、草、普野、种子、爪哇、生长
第 7 类:基因组学	染色体、细胞、核型、染色、畸变、SC(细胞分裂)、群体、着丝点、联会、复合体、着丝粒、动物、多态性、mtDNA、遗传学
第 12 类:动物遗传学	马、多态性、限制性、多样性、黄牛、动物、群体、差异、染色体、mtDNA、物种、血清、起源、地方
第 15 类:植物育种遗传学	水稻、标记、染色体、群体、定位、形状、QTL、品种、杂交、图谱、稻、基因组、突变体、籼稻、抗性、粳稻

表 2　谈家桢学术谱系作者－主题分布

ID	姓名	主题分布概率值 $\times 10^4$				
		第 1 类:血液遗传学	第 4 类:植物遗传学	第 7 类:基因组学	第 12 类:动物遗传学	第 15 类:植物育种遗传学
1	朱立煌	12.4	1 266.4	12.4	12.4	6 122.9
3	曾益滔	5 968.08	24.9	98.6	98.6	12.4
8	冯蜀举	12.4	12.4	9 468.24	9 468.24	24.9

表 3　朱立煌学术谱系作者主题相似度

前十名			后十名		
排名	姓名	主题相似度	排名	姓名	主题相似度
1	尚俊军	0.702	……	……	……
2	韦丽荣	0.688	38	肖晗	0.546
3	毕高峰	0.676	39	赵彬	0.543
4	甘强	0.665	40	王世全	0.542
5	孟征	0.663	41	李仕贵	0.536
……	……	……	42	唐家斌	0.535

3.2　以谈家桢为核心的学术谱系代际间知识传承研究

在生物学理论中，遗传是指子代在连续系统中重复亲代的特性和特征(性状)的现象，其实质是子代承接亲代的遗传物质－基因(决定生物性状)，基因的传递即为遗传。在世代延续过程中，基因的突变使得子代可以发育出非同于亲代的性状，这种现象即为变异[11]。知识被认为是研究过程中由老师(亲代)通过口述、面授等各种方式传递给学生(子代)的基因，笔者借鉴生物学中的遗传和变异的概念，将老师与学生之间以知识为代表的研究方向的延续视之为谱系内知识的“遗传”，将学生在获得知识后在随后的学术生涯中研究方向的转变视为谱系内的知识“变异”。

主题相似度具备连续值域特性，为了方便识别师生知识传承与其学术绩效的相关性，这里借鉴统计学四分位距的知识。四分位距(Interquartile Range，IQR)，又称四分差。是描述统计学中的一种方法，以确定第三、四分位数区间和第一、二分位数区间的区别。与方差、标准差一样，四分位距方法表示统计资料中各变量的分散情形，但四分差更多为一种稳健统计(Robust Statistic)。因此使用四分位距方法对知识传承中的“遗传”与“变异”进行识别和划分具有统计学意义。规定以总样本主题相似度变化区间的前 1/4 值域定义为知识传承中的“变异区间”，主题相似度处在该区间的学者定义为“变异学者”；将总样本主题相似度变化区间的后 1/4 值域定义为知识传承中的“遗传区间”，主题相似度处在该区间的学者定义为“遗传学者”(实际上知识传承中的遗传与变异是相对的，遗传学者存在一定程度的研究主题变异，但变异程度相对较小；变异学者亦存在一定的研究主题遗传，但遗传程度相对较小)。根据上述确定研究主题“遗传”与“变异”的方法，计算出遗传与变异的主题相似度区间，如图 4 所示：



图 4　谈家桢学术谱系遗传与变异的主题相似度

可以看出，朱立煌所延续的学术谱系内有 2 位子代成员存在研究主题遗传现象，有 7 位子代成员发生了研究方向的变异。总体来看，21 位子代成员存在研究主题遗传的现象，有 34 位子代成员发生了研究方向的变异，分别占比 8.97% 和 14.53%。这 54 位学者与其老师的主题相似度如表 4 所示。谱系树包含以谈家桢及其直系学生为核心的共 22 个学术分支，不同分支三类学者的分布数量与比例见图 5。

一般来讲，学生未获得学位之前的学术研究是追随老师脚步的，因此学生毕业之前的研究主题通常与老师一致，变异产生的原因显然来自于学生毕业之后研究主题发生的变化。因此，运用预处理后的第二类样本作为 LDA 主题模型的输入，具体过程如下：①将每一位学者的所有文献根据其毕业时间点与发表时间点的划分为毕业前发表文献集与毕业后发表文献集；

表 4　谈家桢学术谱系中“遗传学者”与“变异学者”信息

学者姓名	导师姓名	主题相似度	遗传/变异	学者姓名	导师姓名	主题相似度	遗传/变异
宁云山	曾义滔	0.747	遗传	程秋应	沈大棱	0.512	变异
陈云弟	曾义滔	0.539	变异	杨扬	施履吉	0.542	变异
黎怀星	傅继梁	0.726	遗传	孙晓平	施履吉	0.537	变异
戴旭明	傅继梁	0.719	遗传	刘祖洞	谈家桢	0.726	遗传
朱海英	傅继梁	0.699	遗传	卢大儒	谈家桢	0.712	遗传
周长文	傅继梁	0.695	遗传	马庆生	谈家桢	0.687	遗传
姚真真	傅继梁	0.531	变异	任大明	谈家桢	0.547	变异
王水良	傅继梁	0.494	变异	柯励生	谈家桢	0.532	变异
陈坚	傅继梁	0.489	变异	罗敏	谈家桢	0.502	变异
房卫平	季道藩	0.499	变异	傅继梁	谈家桢	0.496	变异
黄引	卢大儒	0.719	遗传	汪旭	薛京伦	0.693	遗传
王飞	卢大儒	0.533	变异	冯登敏	薛京伦	0.543	变异
田景琰	罗敏	0.732	遗传	朱作峰	杨金水	0.694	遗传
骆天红	罗敏	0.721	遗传	姜瑞华	杨金水	0.498	变异
张芳琳	罗敏	0.704	遗传	吴益民	赵寿元	0.717	遗传
蔡东升	罗敏	0.551	变异	何以丰	赵寿元	0.704	遗传
尹晓	罗敏	0.507	变异	刘珊	赵寿元	0.694	遗传
韩峻峰	罗敏	0.496	变异	李平	赵寿元	0.531	变异
王晓	罗敏	0.487	变异	尚俊军	朱立煌	0.702	遗传
戴美学	马庆生	0.688	遗传	韦丽荣	朱立煌	0.688	遗传
肖伟	任大明	0.537	变异	梁国华	朱立煌	0.552	变异
董皓林	任大明	0.524	变异	欧阳振乾	朱立煌	0.55	变异
鲍宝龙	任大明	0.519	变异	肖晗	朱立煌	0.546	变异
徐玲	任大明	0.512	变异	赵彬	朱立煌	0.543	变异
蔡蕾	任大明	0.499	变异	王世全	朱立煌	0.542	变异
温建国	沈大棱	0.701	遗传	李仕贵	朱立煌	0.536	变异
徐仁尔	沈大棱	0.545	变异	唐家斌	朱立煌	0.535	变异
孙晓波	沈大棱	0.533	变异				

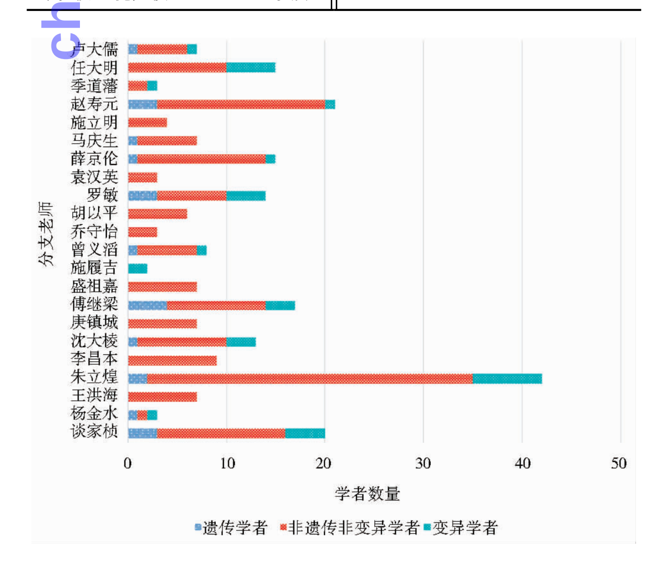


图 5　谈家桢学术谱系各分支三类学者数量统计图

②将学者 i 毕业前文献集与毕业后文献集分别作为两个不同实体 A_i 和 B_i，利用 LDA 模型获取 A_i 和 B_i 所对应的主题分布向量 P_a 和 P_b；③通过计算 P_a 和 P_b 的 JS 距离，确定 A_i 和 B_i 的主题相似度，最后即得到学者 i 毕业前后的主题相似度。如图 6 所示，学生毕业前后主题相似度的变化区间为[0.504,0.780]，遗传与变异的主题相似度区间分别为[0.711,0.780]和[0.504,0.573]。



图 6　谈家桢遗传学学术谱系遗传与变异的主题相似度区间划分

同理计算出学者毕业前后研究主题发生遗传与变异的比例分别为 9.66% 和 24.14%。“遗传学者”有 14 人，“变异学者”有 35 人。很明显，与代际间学术传承所得的研究主题遗传变异结果相比，在学生自身研究主题毕业前后变化层面上，“遗传学者”比例基本持平，而“变异学者”的比例有明显的提升。

4　以谈家桢为核心的学术谱系内研究主题相似度与学术绩效的相关性分析

本研究试图寻找出以文本内在联系为代表的知识传承是否对学者的职业生涯有影响，如有影响，知识传承在何种程度以何种方式影响人才的职业生涯发展？为此，笔者对学术谱系内成员的知识传承程度与其个人成长的科研绩效进行了相关性分析，这里以 h 指数作为衡量学术谱系成员的科研绩效评价指标。

如图 7 中所示，图 7(1)和图 7(2)展示的是以谈家桢为核心的学术谱系内学者代际间主题相似度与其 h 指数的分布图，图 7(3)和图 7(4)展示的是学术谱系内学者毕业前后主题相似度与其 h 指数的分布图。为了方便分析研究结果，将学者代际间研究主题相似度与相应的 h 指数数据称为组合 1，学者毕业前后研究主题相似度与相应的 h 指数数据称为组合 2。图中不同颜色区域分别代表“变异学者”(Variation)、“非遗传非变异学者”(Non-hereditary Non-variation)、“遗传学者”(Hereditary)。

图 7(1)与图 7(3)两图分别是组合 1 与组合 2 数据的散点分布，可以看出，图 7(1)中点的分布具有聚集特性，71% 的学者的相似度落在(0.55,0.65)的狭窄区间内；“遗传学者”“非遗传非变异学者”“变异学者”人数所占比例分别为 8.97%、76.50%、14.53%，h 指数均值分别为 5.42、4.54、7.10，可以看出“遗传学者”“变异学者”人数虽然占比较少，但相对于“非遗传

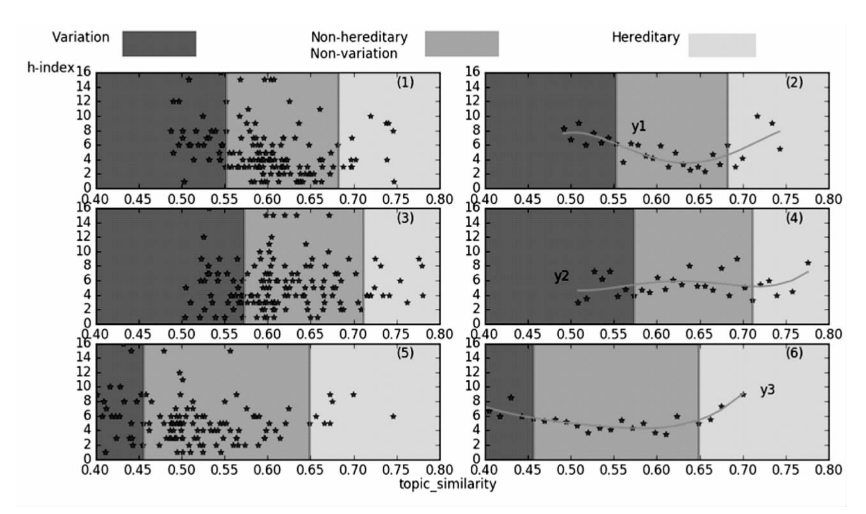


图 7 学者主题相似度与 h 指数分布图

非变异学者”具有更高的学术绩效；h 指数在 5 以上的高绩效作者[12]在三个区间内的比例分别为 3.00%、56.92%、35.38%，可以看出“变异学者”中高绩效占比相对更高。同样，在图 7(2) 中“遗传学者”“非遗传非变异学者”“变异学者”人数所占比例分别为 9.66%、66.20%、24.14%，h 指数均值分别为 5.42、5.52、5.11；高绩效作者比例分别为 8.00%、76.00%、16.00%，可以看出，“遗传学者”与“变异学者”在高绩效科研学者群体中所占比例明显高于二者在谱系内总体学者群体中所占比例。另外，为了进一步寻找学者主题相似度与其 h 指数之间是否存在相关性，将所有学者根据其主题相似度区间划分至 30 等份，在每个小区间上取一个点，该点的横轴为该区间所对应的 h 指数的均值，纵轴值取区间中间点对应的数值。依照该方法，分别对组 1 与组 2 两组数据进行处理后得到散点图，如图 7(2) 和图 7(4) 所示，进一步对散点图进行曲线拟合，图中蓝色实线是对散点图进行四次多项式拟合后的结果，四次多项式分别为：

$$y_1 = -12823.5x^4 + 18522.5x^3 - 9533.4x^2 + 2050.1x - 147.8$$

$$y_2 = 4934.2x^4 - 7555.2x^3 + 4201.7x^2 - 1005.8x + 93.0$$

y_1 拟合后的 $R^2 = 0.699$；y_2 拟合后的 $R^2 = 0.068$。

评价最小二乘法进行的曲线拟合优度的标准是 R^2，R^2 值越大，拟合效果越好，R^2 越小，拟合效果越差。R^2 取值介于 0 到 1 之间。显然，y_1 拟合效果较好。从图 7(2) 可以看出，曲线拟合后的分布是扁平凹型分布，即中间低，两边高。凹型分布代表着代际间主题相似度处于“遗传”与“变异”区间内的学者的学术绩效要明显高于“非遗传且非变异”的学者，表明科研绩效较高的人群不是人数最多的“非遗传与非变异”学者，而恰恰正是处于少数群体的遗传学者和变异学者。从图 7(4) 中可以看出，拟合后 h 指数的均值在学生毕业前后主题相似度的变化不明显，始终保持在 4–8 的水平。

上述代际间的研究主题相似度揭示了学生整体研究主题和老师研究主题的变化特征，未能回答学生毕业以后研究方向与导师是否存在差异的问题。为此，笔者还进一步对学生毕业以后的研究主题和老师的研究主题进行了相似度计算。首先选取学生毕业后的文献集与老师文献集进行主题抽取，并计算两者之间的主题相似度，进而将主题相似度与 h 指数进行相关性分析，研究结果如图 7(5) 和图 7(6) 所示。从图中可以看出，学生毕业以后研究主题与老师研究的相似度区间为(0.27,0.74)，相似度取值范围增加，这是由于移除了两实体对照组的重复文本（师生在研究生期间共同发表的文章）影响；“遗传学者”“非遗传非变异学者”“变异学者”所占比例分别为 6.20%、68.21%、25.58%，h 指数均值分别为 6.62、4.81、7.24，可以看出“遗传学者”和“变异学者”的学术绩效要明显高于“非遗传非变异学者”；h 指数在 5 以上的高绩效作者在三个区间内所占比例分别为 10.25%、57.69%、32.05%，可以看出与组合 1 结论相似，“遗传学者”与“变异学者”在高绩效科研群体中所占比例相较于二者在谱系内总体学者群体中所占比例有所提高；拟合曲线同样具有扁平凹型分布，再次证实了知识“遗传”与“变异”有助于学术绩效提高的结论，其中：

$$y_3 = -3762.8x^4 + 6857.6x^3 - 4464.7x^2 + 1223.1x - 112.73$$

y_3 拟合后的 R^2 值为 0.70。

5 谈家桢学术谱系内遗传与变异学者的研究主题分布

上述研究显示，“遗传学者”与“变异学者”在其职业生涯发展过程中取得了相对较高的学术绩效。同时，笔者对与这两种类型学者的研究主题也产生了浓

厚的兴趣，他们的研究主题是何种分布？是否存在显著差别？针对上述疑问，进一步利用主题模型方法对“遗传学者”和“变异学者”的研究主题分布进行了具体分析。

LDA主题模型获得的θ的最大似然估计值表示每个学者的主题分布，把学者文献集扩展到谱系内的学者群，同样可以求出该群体语料的主题分布 θ 。设 θ_k^d 为学者d的主题k所占的比例，则学者群s在主题k的强度 θ_k^s 如公式(6)所示，其值介于0到1之间[13]。

$$\overline{\theta_k^s}=\frac{\sum_{d=1}^{M}\theta_k^d}{M} \qquad \text{公式(6)}$$

用每个主题的 $\overline{\theta_k^s}$ 表示目标学者群体s中主题k的强度，从而可以得出目标谱系人群的主题分布，如图8所示。其中，蓝色代表谈家桢学术谱系内根据代际间主题相似度计算得出的“遗传学者”。同理，绿色代表“变异学者”，红色代表学术谱系全体学者群体。从图中可以看出：①上述三个群体主题强度最大的都是主题14（细胞遗传学），“遗传学者”“变异学者”“全体学者”在该主题的强度分别为0.53、0.27、0.33，经过咨询可知该研究主题是遗传学的基础性研究，是进行其他研究的前提；②“遗传学者”的研究主题主要分布在主题12（动物遗传学）、主题14（细胞遗传学）、主题17（肿瘤细胞遗传学）、主题20（群体遗传学），其他主题分布很少；③“总体学者”与“遗传学者”的研究主题分布较为分散，分布相对“遗传学者”较为均匀；④部分“变异学者”致力于植物基因组研究、医学遗传学研究、抗毒领域应用研究，而“遗传学者”在上述领域没有涉足；⑤部分“遗传学者”致力于动物遗传学的研究，而“变异学者”在该领域几乎很少有研究。

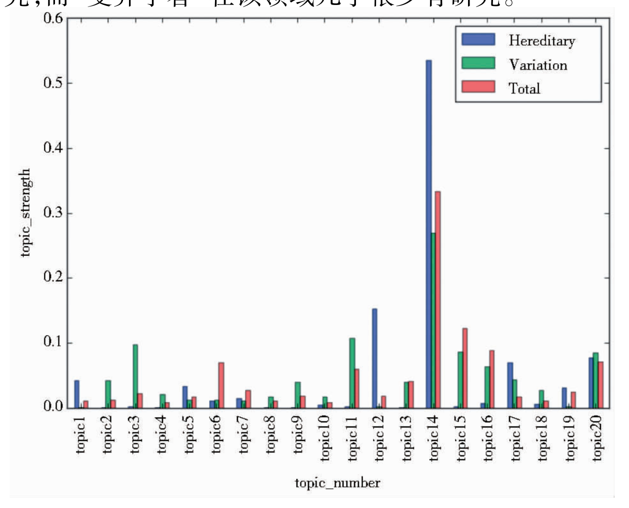


图8　学者群体主题分布图

6　结论与展望

笔者运用LDA主题模型研究了以谈家桢为核心的遗传学学术谱系内的主题分布，在语义层面探索了该谱系成员的研究主题分布，通过计算学者研究主题的相似度将学者划分为“遗传学者”“非遗传非变异学者”“变异学者”三种类型，进一步将主题相似度与学者的科研绩效进行相关性分析，研究结果表明：①谱系内成员的科研绩效与代际间研究主题的变化程度具有相关性，与学者毕业前后研究主题的变化度无关；②“遗传学者”与“变异学者”的平均科研绩效要明显高于“非遗传与非变异学者”的平均科研绩效，即一定程度上的“尊承前贤”与“开疆拓土”有利于提升学者的科研绩效；③“遗传学者”和“变异学者”的研究主题分布有较大差异，总体来看，细胞遗传学依然是遗传学领域中的基础研究领域，三类学者的研究都需要该领域知识积累的支撑。

本研究有效地在语义层面提取了学者研究主题，并通过JS距离测度了谈家桢学术谱系内代际间的主题相似度，相较于以往定性研究提高了研究的可信度。基于研究主题的人才发展路径研究有助于人才成长规律的揭示和人才政策的制定，科研评价体系和科研激励机制应该充分考虑知识传承的“遗传基因”，给予继续坚守在同一个研究方向的科研人员以较长时间的研究周期，以促进其在扎实的研究基础上进行深入的科学创新；另一方面鼓励研究人员在知识传承的基础上涉猎新兴领域或交叉领域，给予资金资助或资源扶持，帮助其在知识转型领域取得更大进步。

本文的研究仅限于对遗传学领域学术谱系知识传承的研究，研究结果具有一定的领域局限性，今后将考虑学科的差异性，进一步对其他研究领域学术谱系的知识传承进行探索和研究。同时，中外科学研究领域的特征差异是显而易见的，未来将进一步对国外学术谱系知识传承的特征进行深入考察和研究。

参考文献：

[1] 刘颖，张燕蕾，张大庆.中国科学家学术谱系库的构建思路初探与实践[J].图书情报工作，2014,58(S2):60－62.

[2] CRONIN B，SUGIMOTO C. Academic genealogy[C]//BLAISE C,CASSIDY R. Beyond bibliometrics. London：MIT Press，2014：480.

[3] JACKSON D C. Academic genelogy and direct calorimetry：a personal account[J]. Advances in physiology education，2011，35(2),120－128.

[4] MALMGREN R D, OTTINO J M, AMARAL L A N. The role of mentorship in protégé performance[J]. Nature, 2010, 465(7298):622-626.
[5] 常欢，吕瑞花，张佳静. 学术谱系内合作网络研究——以刘东生为核心的第四纪学术谱系为例[J]. 情报理论与实践，2016, 39(4):14-19.
[6] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation[J]. Journal of machine learning research, 2003, 3(3):993-1022.
[7] ROSEN Z, MICHA L, GRIFFITH S, et al. The author-topic model for authors and documents[C]//Proceedings of the 20th conference on Uncertainty in artificial intelligence. New York: AUAI, 2004: 487-494.
[8] DHILLON I S, MODHA D S. Concept decompositions for large sparse text data using clustering[J]. Machine learning, 2001, 42(1):143-175.
[9] 史庆伟，乔晓东，徐硕，等. 作者主题演化模型及其在研究兴趣演化分析中的应用[J]. 情报学报，2013, 32(9):912-919.
[10] WANG X, MCCALLUM A. Topics over time: a non-Markov continuous-time model of topical trends[C]// BACKSTROM L, HUTTENLOCHER D, KLEINBER G, et al. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. New York: ACM, 2006:424-433.
[11] 谢平. 生命的起源-进化理论之扬弃与革新[M]. 北京:科学出版社，2014.
[12] 刘俊婉，郑晓敏，宿娜，等. 国内外情报学领域期刊发文时滞的计量分析——以 Scientometrics 和《情报学报》期刊为例[J]. 中国科技期刊研究，2016, 27(12):1292-1299.
[13] 崔凯. 基于 LDA 的主题演化研究与实现[D]. 长沙:国防科学技术大学，2010.

作者贡献说明：

刘俊婉:研究命题的提出与设计,数据分析,论文撰写;
杨波:数据采集、清洗,程序代码设计,数据分析与论文起草;
王菲菲:论文设计与论文修改;
徐硕:主题模型方法指导。

Research on Knowledge Inheritance of Academic Pedigree Based on LDA Topic Model
——A Case Study of Genetics Pedigree with the Core of Tan Jiazhen

Liu Junwan　Yang Bo　Wang Feifei　Xu Shuo
College of Economic and Management, Beijing University of Technology, Beijing 100124

Abstract: [**Purpose/significance**] Academic pedigree promotes science development by the way of knowledge inheritance. It is of great reference value to study the characteristics of knowledge transmission and explore the effect of inheritance model on academic output, and it is of great reference value for the relevant departments to find out the law of talent growth and formulate scientific and technological personnel training policy. [**Method/process**] By the method of LDA topic model, this paper took the journal literature of genetics published in CNKI database as research object, and quoted the concept of "hereditary" and "variation" in biology. Then, according to the topic similarity, we divided pedigree members into "hereditary scholars", "variation scholars" and "non-hereditary non-variation scholars", and analyzed the academic performance of these three kinds of scholars. [**Result/conclusion**] The results show that the academic performance of "hereditary scholars" and "variation scholars" in the academic pedigree of Tan Jiazhen is relatively high; The number of "non-hereditary non-variation scholars" is the largest, but their academic performance is relatively low; For different topics, the distribution of "variation scholars" and "hereditary scholars" is significantly different.

Keywords: academic pedigree　knowledge inheritance　topic model　genetics